Final Project

Tanner Oates

2/23/2021

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)
hsb2 <- read.table("../data/hsb2.txt", header=TRUE, sep="\t")
```

Introduction:

The data set I will be working with is from a study called High School and Beyond. The data set includes data from students, parents, teachers, school administrators, and administrative records. The study followed the development of young people into adulthood. My variables of interest are school type, science, and math. I want to see if people who attended private school have a better science/math understanding than those who are in public school.

Univariate description:

Variable name: ‘schtyp’, ‘science’, and ‘math’

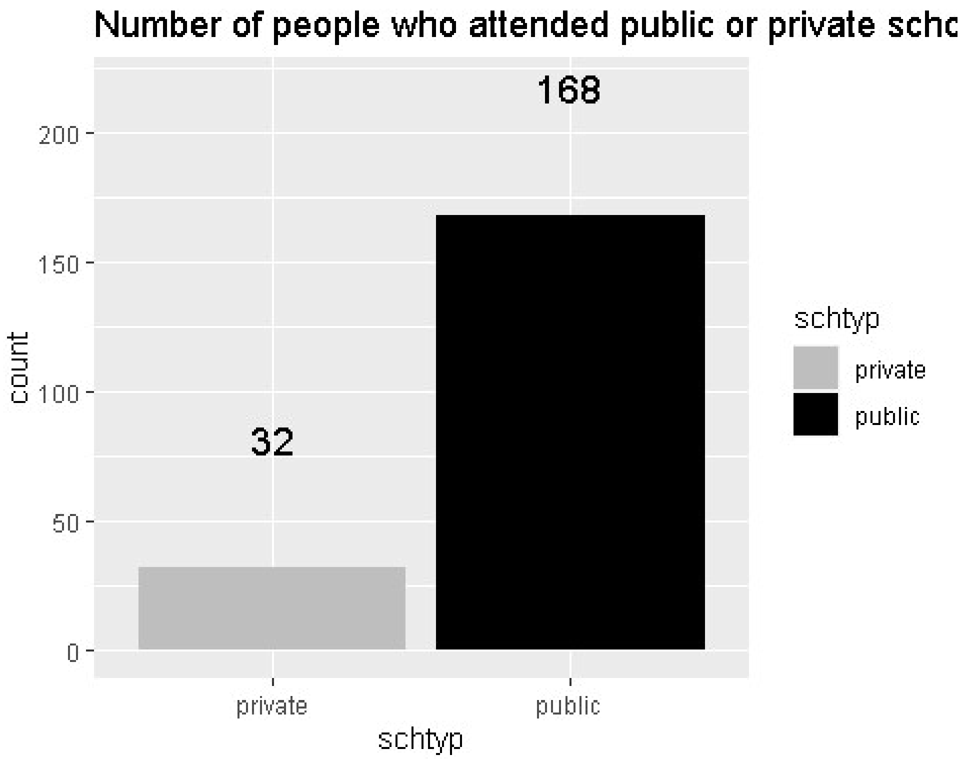
School Type

```
table(hsb2$schtyp)

##
## private  public
##      32     168
```

This table shows the number of public and private school attendees.

```
ggplot(hsb2, aes(x=schtyp, fill=schtyp)) +
    geom_bar() + ggtitle("Number of people who attended public or private
school") +
    geom_text(aes(y=..count.. + 50, label=..count..), stat='count', size = 5)
+ scale_fill_manual(values=c("gray","black"))
```



This bar graph shows that there are 32 private school attendees and 168 public school attendees. The data also tells us that the study has a wider variety of outcomes for public school attendees than private.

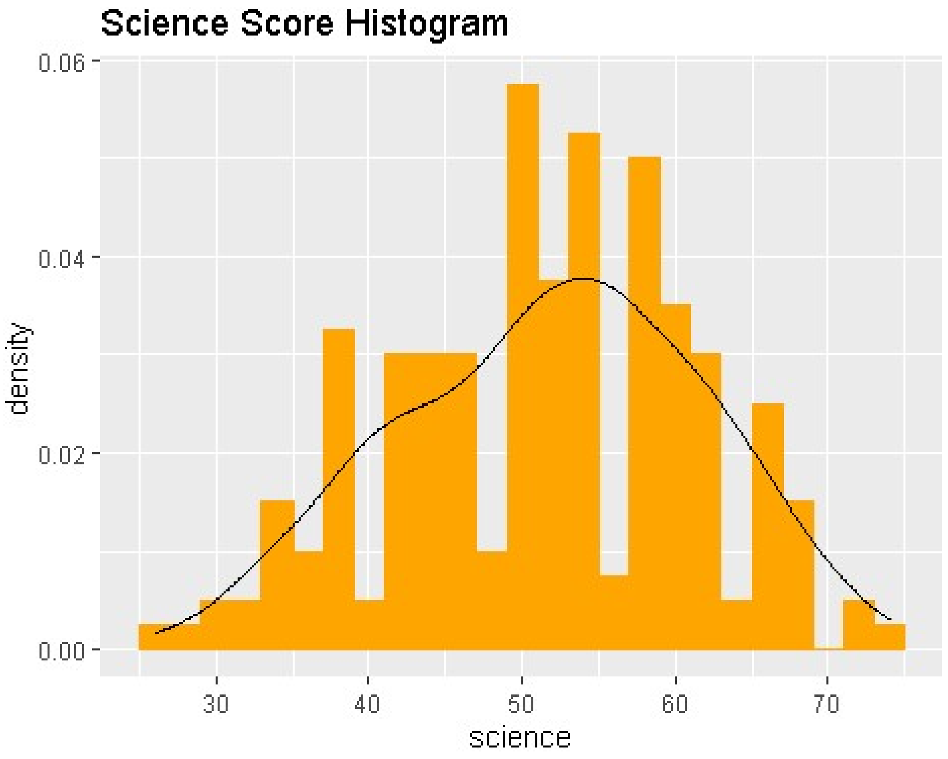
Science Variable

```
summary(hsb2$science)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   44.00   53.00   51.85   58.00   74.00
```

The maximum score for science is 74 and the minimum was 26. The mean and median were fairly close being 51.85 and 53. I would like to know more about why the max and min are so far off from each other. Maybe the higher score could be coming from someone in the science field.

```
ggplot(hsb2, aes(x=science)) + geom_histogram(aes(y = ..density..), col =
"orange", fill = "orange", bins= 25) + geom_density(col="black")+
ggtitle("Science Score Histogram")
```



This is a histogram of the science scores. The graph shows a concentration of scores around 50. Overall though the graph shows a typical bell curve you would expect to see. The overlaid density plot illustrates the shape of the graph.

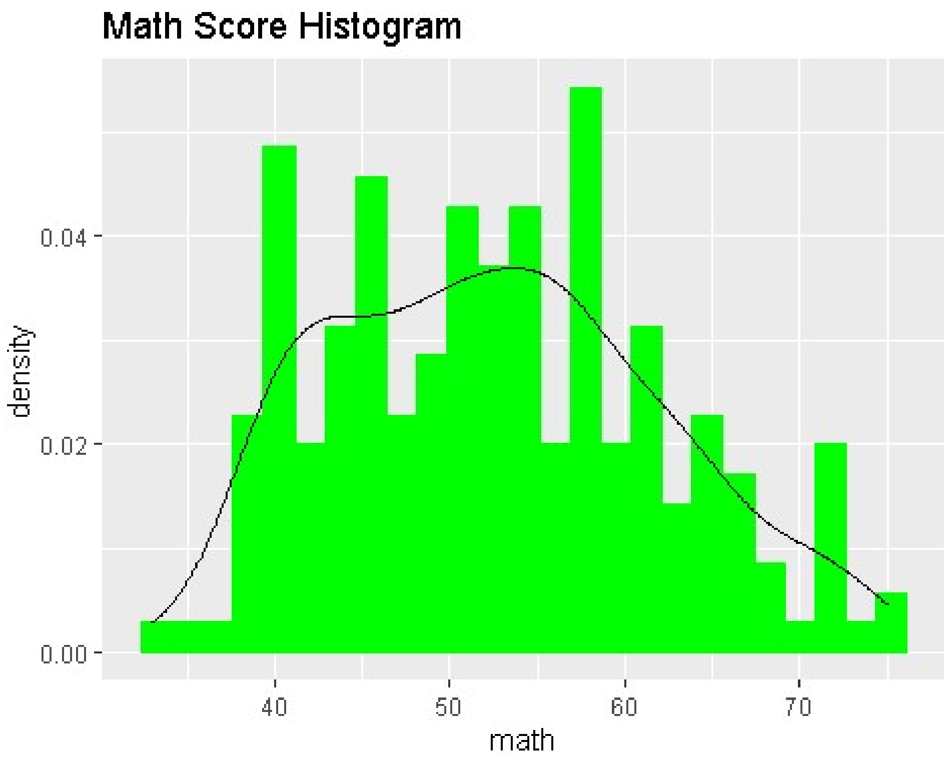
Math Variable

```
summary(hsb2$math)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   33.00   45.00   52.00   52.65   59.00   75.00
```

The maximum score for math is 75 and the minimum was 33. The mean and median were close being 52.65 and 52. This is interesting as the scores are almost identical to that of the science scores.

```
ggplot(hsb2, aes(x=math)) + geom_histogram(aes(y = ..density..), col =
"green", fill = "green", bins= 25) + geom_density(col="black")+ ggtitle("Math
Score Histogram")
```

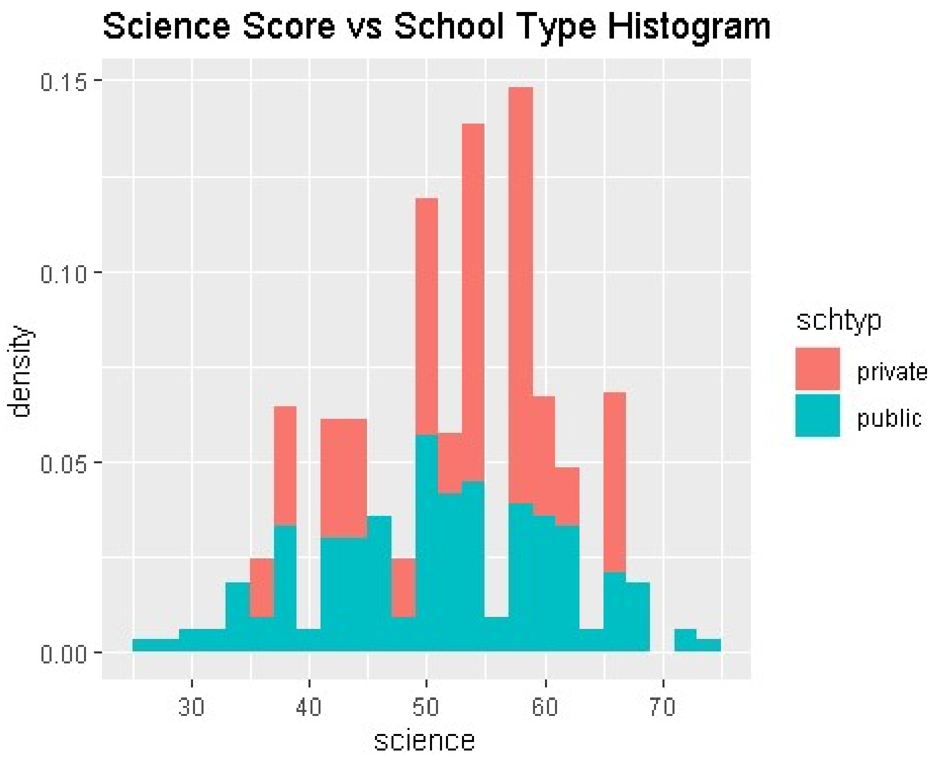


This is a histogram of the math scores. Most math scores were in the 40-60 range. The histogram is very similar to the one of the science scores although the density plot is slightly different.

Bivariate comparison

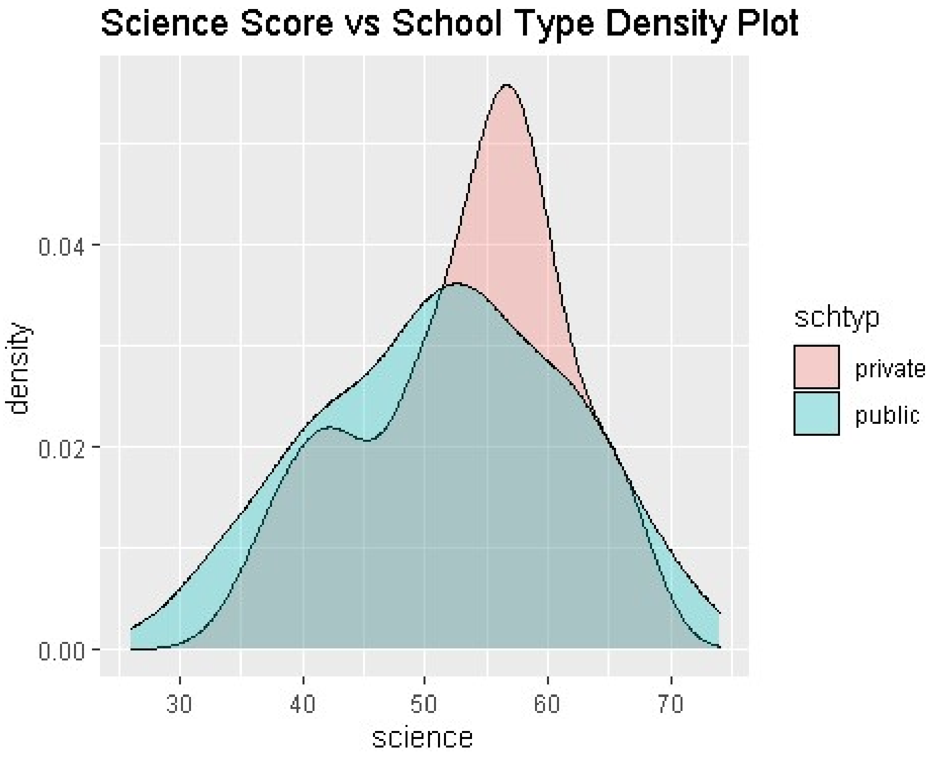
Science scores vs. School Type

```
ggplot(hsb2, aes(x=science, fill = schtyp)) + geom_histogram(aes(y =
..density..), bins= 25) + ggtitle("Science Score vs School Type Histogram")
```



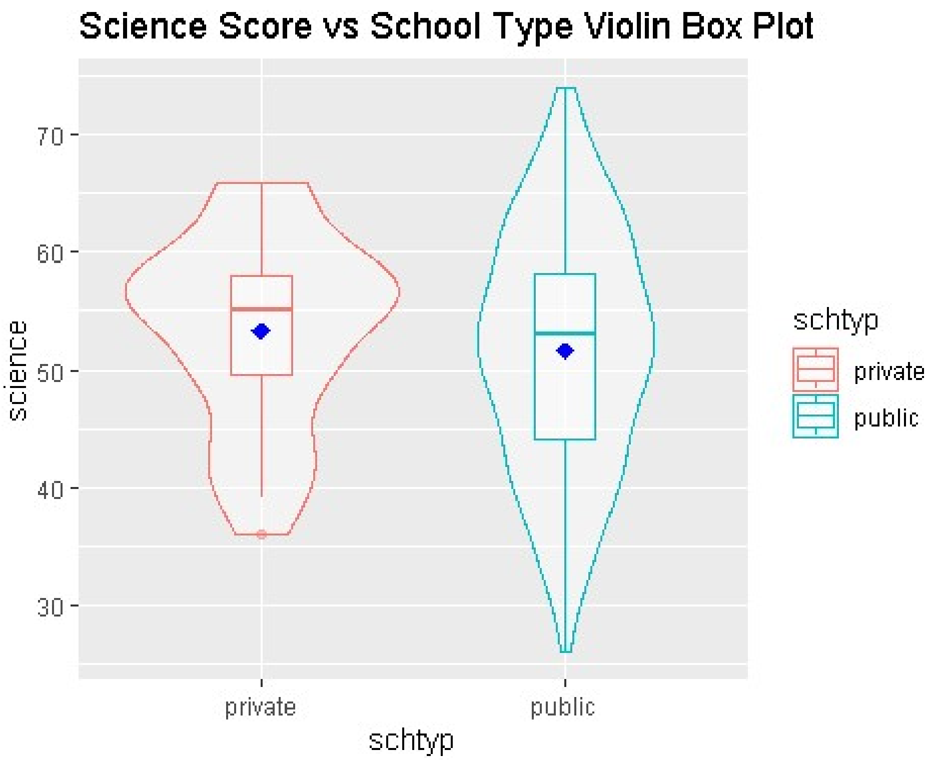
This histogram shows the science scores of both private and public school attendees. The pink sections are private school scores and the cyan is the public school scores. The graph shows that while public schools did have the highest scores they also had the lowest. But whats more interesting is the graph clearly shows that people who went to private school scored better in science.

```
ggplot(hsb2,    aes(x   = science,  fill    = schtyp))  +
geom_density(alpha=0.3) + ggtitle("Science Score vs School Type Density
Plot")
```



This density plot is another way of showing how the scores are distributed. It also gives us a second look at how the private school scores trended on the higher side.

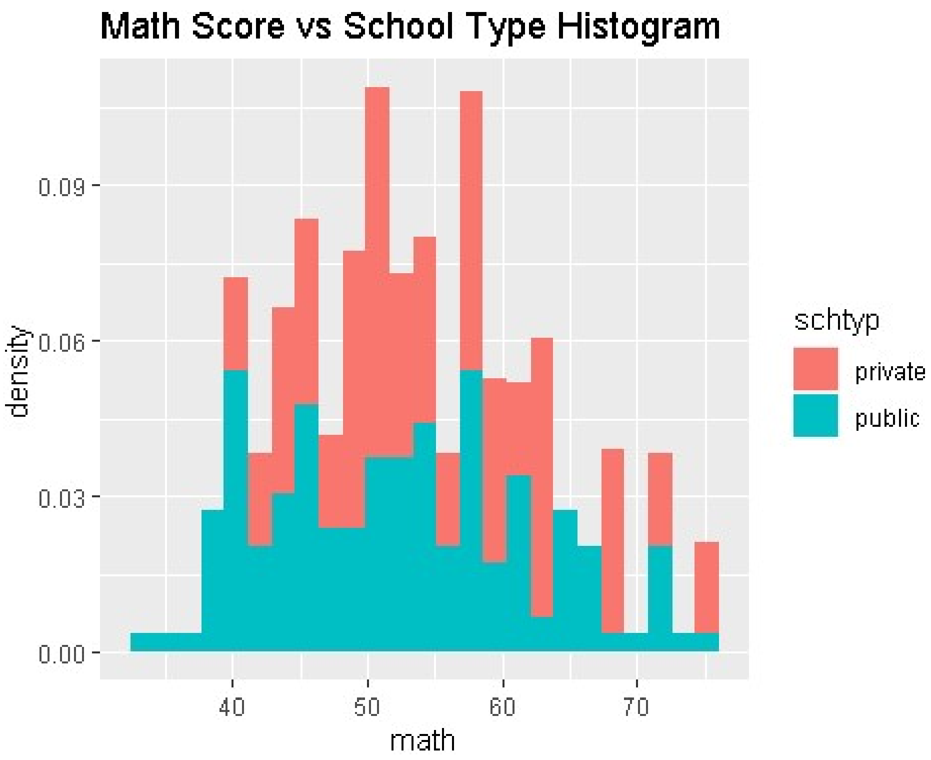
```
ggplot(hsb2, aes(x=schtyp,  y=science,  col=schtyp)) + geom_violin(alpha=.4)
+ geom_boxplot(alpha=.4,   width=.2) + stat_summary(fun= "mean", geom=
"point", size=3, pch= 18, color="blue") + ggtitle("Science Score vs School
Type Violin Box Plot")
```



By looking at the box plot we see again that the private school scores are more condensed having a far smaller range than public school scores. The blue dots in the middle of the plots represent the mean of the scores for each type of school.

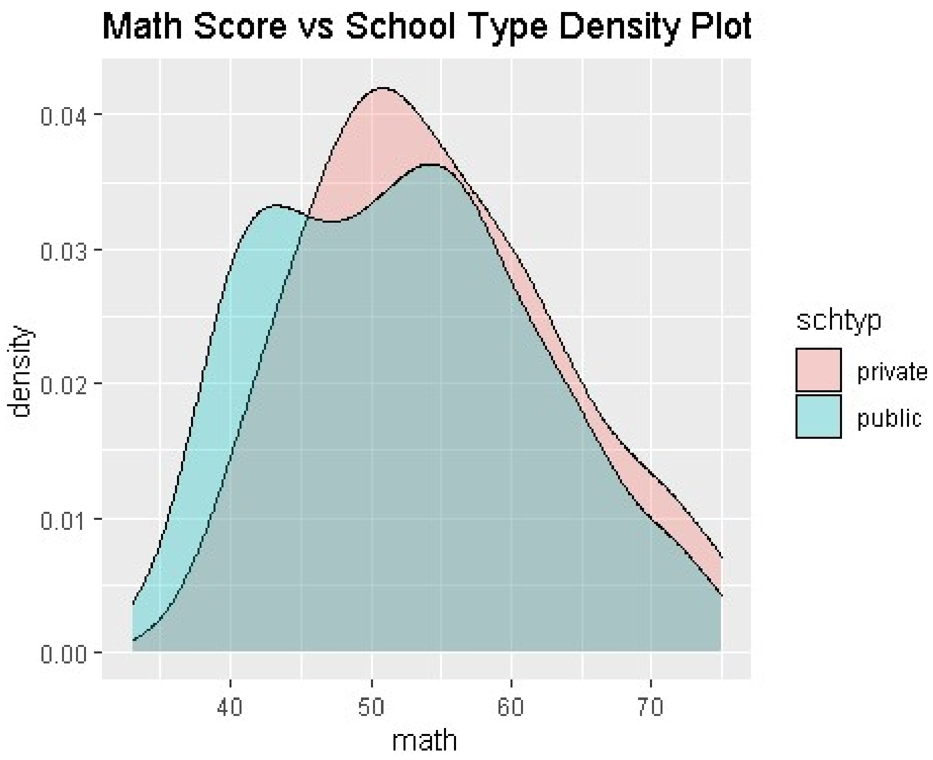
Math scores vs. School Type

```
ggplot(hsb2, aes(x=math, fill = schtyp)) + geom_histogram(aes(y =
..density..), bins= 25) + ggtitle("Math Score vs School Type Histogram")
```



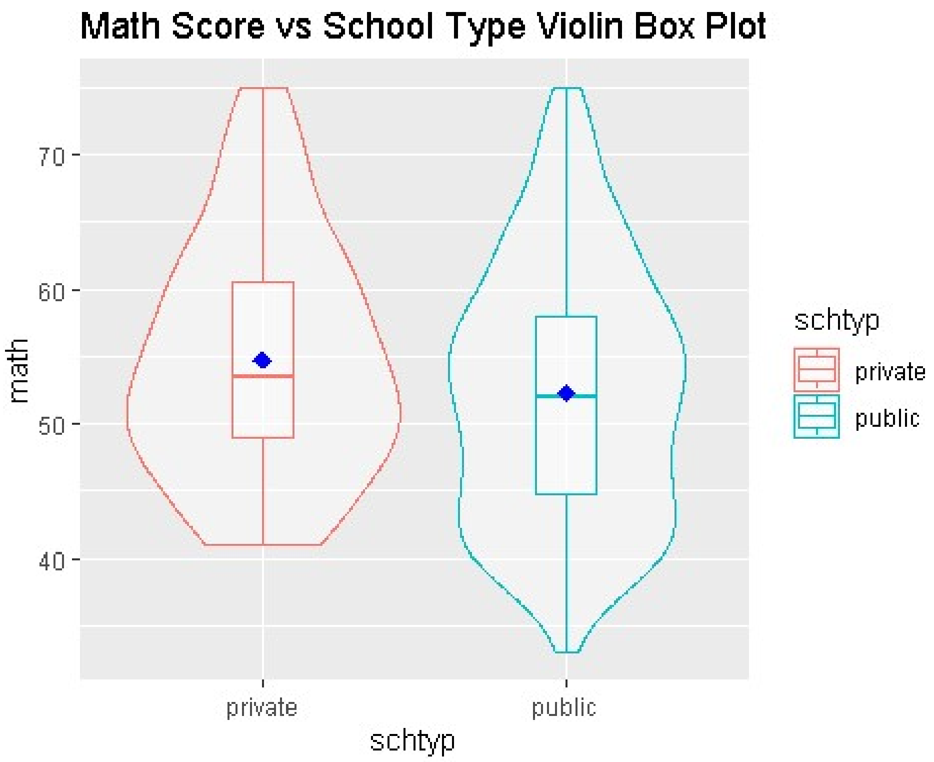
This histogram shows the math scores of both private and public school attendees. The pink sections are private school scores and the cyan is the public school scores. The graph shows that private school attendees scored higher than public school attendees but its not a very big difference.

```
ggplot(hsb2,    aes(x   = math, fill    = schtyp))  + geom_density(alpha=0.3)
+ ggtitle("Math Score vs School Type Density Plot")
```



This density plot is another way of showing how the scores are distributed. It also gives us a second look at how the private school scores trended on the higher side. From looking at the graph you can really see why the mean for math scores is 52.65.

```
ggplot(hsb2, aes(x=schtyp,  y=math, col=schtyp)) + geom_violin(alpha=.4) +
geom_boxplot(alpha=.4,   width=.2) + stat_summary(fun= "mean", geom= "point",
size=3, pch= 18, color="blue") + ggtitle("Math Score vs School Type Violin
Box Plot")
```



This box plot shows that while private school attendees mean is higher, the top half of both plots are about the same. The main difference between the two is public school has some people with lower scores but they also have a higher population in the data set. It would be interesting to see what the graph would look like if there was a more balanced sample of people.

Conclusion

In conclusion, the data shows that people who attended private school over public school did have on average better scores in science and math. I can't say for sure that school type is a big factor as the number of people who attended private school versus public school is far less, it would better to do these tests with a more balanced sample of people.